Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network

Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang,
and Kwang-Ting Cheng, *Fellow, IEEE*

***Abstract*—Automated methods for detecting clinically significant (CS) prostate cancer (PCa) in multi-parameter magnetic resonance images (mp-MRI) are of high demand. Existing methods typically employ several separate steps, each of which is optimized individually without considering the error tolerance of other steps. As a result, they could either involve unnecessary computational cost or suffer from errors accumulated over steps. In this paper, we present an automated CS PCa detection system, where all steps are optimized jointly in an end-to-end trainable deep neural network. The proposed neural network consists of concatenated subnets: 1) a novel tissue deformation network (TDN) for automated prostate detection and multimodal registration and 2) a dual-path convolutional neural network (CNN) for CS PCa detection. Three types of loss functions, i.e., classification loss, inconsistency loss, and overlap loss, are employed for optimizing all parameters of the proposed TDN and CNN. In the training phase, the two nets mutually affect each other and effectively guide registration and extraction of representative CS PCa-relevant features to achieve results with sufficient accuracy. The entire network is trained in a weakly supervised manner by providing only image-level annotations (i.e., presence/absence of PCa) without exact priors of lesions' locations. Compared with most existing systems which require supervised labels, e.g., manual delineation of PCa lesions, it is much more convenient for clinical usage. Comprehensive evaluation based on fivefold cross validation using 360 patient data demonstrates that our system achieves a high accuracy for CS PCa detection, i.e., a sensitivity of 0.6374 and 0.8978 at 0.1 and 1 false positives per normal/benign patient.**

***Index Terms*—CS PCa detection, joint optimization, multimodal registration, neural network.**

Manuscript received October 19, 2017; revised December 21, 2017 and December 27, 2017; accepted December 28, 2017. Date of publication January 3, 2018; date of current version May 1, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61502188, in part by the Wuhan Science and Technology Bureau under Award 2017010201010111, and in part by the Program for HUST Acadamic Frontier Youth Team. *(Corresponding author: Xin Yang.)*

Z. Wang, C. Liu, and X. Yang are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xinyang2014@hust.edu.cn).

D. Cheng is with the University of Bridgeport, CT 06604 USA.

L. Wang is with the Department of Radiology, Tongji Hospital, Huazhong University of Science and Technology, Wuhan 430030, China.

K.-T. Cheng is with the School of Engineering, Hong Kong University of Science and Technology, Hong Kong.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2017.2789181

I. Introduction

Prostate cancer (PCa) is the most commonly diagnosed cancer other than skin cancer, and also one of the leading causes of cancer death among men [1]. Fortunately, 90% of all PCa cases are low-risk whose Gleason score (GS) is equal to or smaller than 6. Low-risk PCa could remain dormant for decades, and thus only needs active surveillance. However, men with clinically significant (CS) PCa whose GS is equal to or greater than 7 could experience high fatality rates and the mortality increases year by year if there is no timely diagnosis and proper treatment [2]. Therefore, accurate detection of CS PCa is crucial to avoid over-treatment of low-risk tumors and to reduce mortality. Multi-parameter magnetic resonance imaging (mp-MRI), which typically includes T2-weighted (T2w) imaging, diffusion-weighted imaging (DWI), and dynamic contrast-enhanced (DCE) MRI, captures both anatomical and functional information of a prostate and thus is becoming increasingly popular for CS PCa detection [3], [4]. However, interpreting mp-MRI data is a very demanding task which requires substantial expertise and significant labor from radiologists. Therefore, there is a strong need for automated detection of CS PCa in mp-MRI images to alleviate such demanding requirements in reading and to reduce risk of over-/under-treatment.

In the past decade, several solutions for computer-aided PCa detection and diagnosis (CAD) in mp-MRI images have been proposed [3], [5]–[13]. In general, existing CAD systems typically consist of three separate steps [14]: i) image registration to remove misalignments among different modalities due to inevitable patients' motion and/or different acquisition parameters, ii) prostate segmentation which helps guide the focus of the subsequent processes only on prostate regions, iii) candidate lesion generation and classification to detect suspicious lesions and differentiate CS from nonCS tumors. Several methods have been developed for each step. For multimodal registration, the authors in Peng *et al.* [4], Artan *et al.* [5], and Fehr *et al.* [6] rectified the misalignment across different image modalities manually according to several anatomical landmarks, such as urethra, ejaculatory ducts, prostatic capsule, etc. Other works [3], [15]–[17] employed automated registration methods which typically search for an optimized geometric transformation that can maximize the

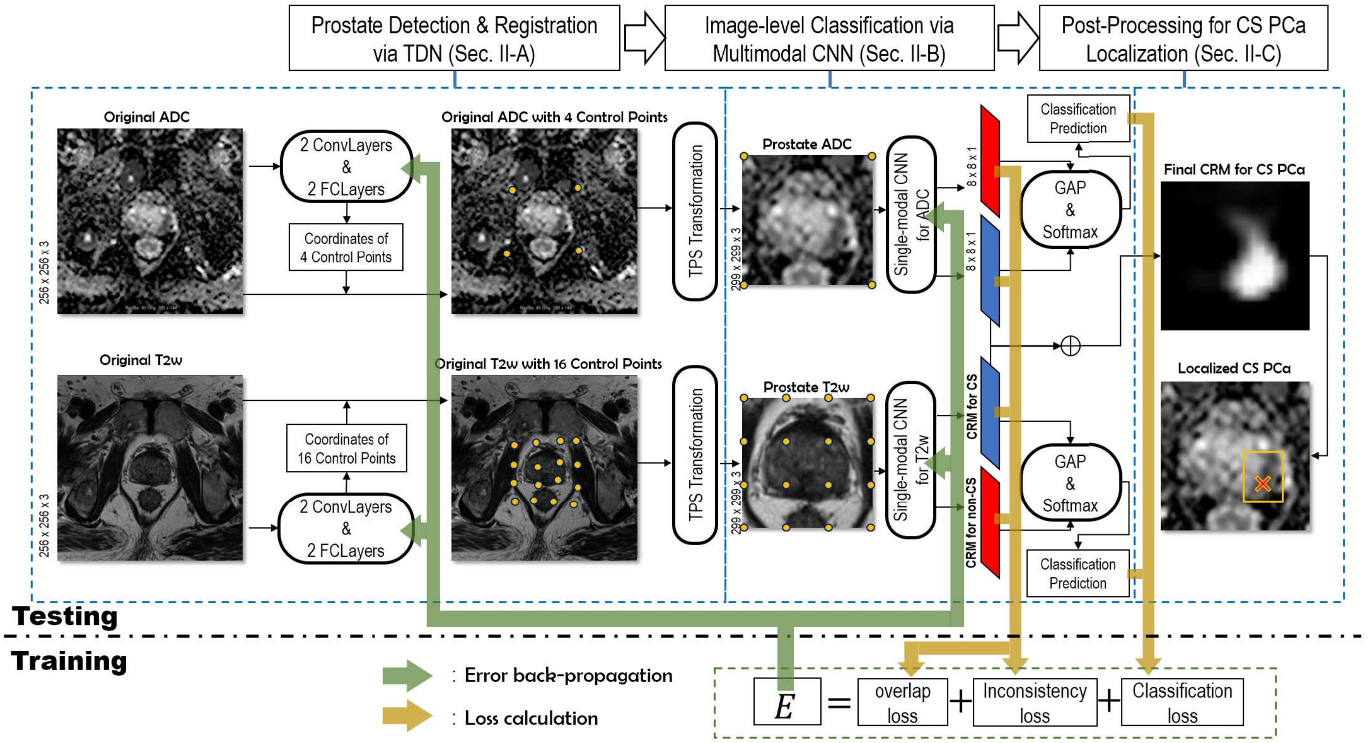
Prostate Detection & Registration via TDN (Sec. II-A)
Image-level Classification via Multimodal CNN (Sec. II-B)
Post-Processing for CS PCa Localization (Sec. II-C)
Original ADC
256 x 256 x 3
2 ConvLayers & 2 FCLayers
Coordinates of 4 Control Points
Original ADC with 4 Control Points
TPS Transformation
Prostate ADC
299 x 299 x 3
Single-modal CNN for ADC
8 x 8 x 1
8 x 8 x 1
Classification Prediction
GAP & Softmax
Final CRM for CS PCa
Original T2w
256 x 256 x 3
Coordinates of 16 Control Points
2 ConvLayers & 2 FCLayers
Original T2w with 16 Control Points
TPS Transformation
Prostate T2w
299 x 299 x 3
Single-modal CNN for T2w
CRM for CS
CRM for non-CS
GAP & Softmax
Classification Prediction
Localized CS PCa
Testing
Training
: Error back-propagation
: Loss calculation
E = overlap loss + Inconsistency loss + Classification loss

Fig. 1. The framework of our unified system for jointly prostate detection, ADC-T2w registration and CS PCa localization.

mutual information (MI) [18] between images of different modalities. For prostate segmentation, several studies [5], [11], [19]–[24] rely on manual delineation of the prostate gland's boundary in every slice, while other works developed automated methods to minimize the manual efforts in this task. For example, Viswanath *et al.* [25] proposed to use an active shape model based on Multi-Attribute Non-initializing Texture Reconstruction (MANTRA) for prostate segmentation and in [3] and [8] Litjens *et al.* used multi-atlas-based segmentation [26]. For PCa detection and diagnosis, a number of feature representation and classification methods have been proposed [3], [4], [6], [8], [12], [27], [28]. Tiwari *et al.* [28] employed features from MRSI and T2w to identify PCa-affected voxels. Litjens *et al.* [8] represented each voxel using intensities and blobness of ADC, homogeneity, etc. This work was further enhanced in [3] by integrating both anatomical and pharmacokinetic voxel features. Peng *et al.* [4] distinguished cancer foci from normal foci based on features including the 10th percentile and the average ADC values, T2w intensity histogram skewness and Tofts K^{trans}. Fehr *et al.* [6] represented regions using first- and second-order texture features computed from the T2w and ADC images. Vos *et al.* [29] detected blobs in ADC images as candidates and then filtered outliers using a shape prior.

Despite the success of existing methods, they rely on separate steps which are optimized individually without considering performance of the other steps. As a result, such methods would either spend some unnecessary computation in a single step because of ignoring error tolerance ability of the other steps, or suffer from significant errors accumulated over stages. Additionally, many existing methods demand non-trivial manual operations and/or parameter settings for each step, yielding inconvenience and high cost of human resources and time in practical clinical usage. Moreover, existing methods mainly employ handcrafted features and simple multimodal feature fusion solutions, e.g. feature concatenation [30] and weighted summation of results [31]. The feature robustness and distinctiveness, the relevance among multimodal features and the methods for effectively fusing multimodal information from mp-MRI have not been well studied.

To address the above limitations, in this work we for the first time propose an automated CS PCa detection system which jointly optimizes all steps, i.e. prostate detection, multimodal registration and CS PCa detection, in an end-to-end trainable deep neural network. Specifically, the deep neural network consists of two sub-networks: 1) a tissue deformation network (TDN) which automatically estimates a thin plate spline (TPS) transformation for joint prostate region detection and registration on pairs of Apparent Diffusion Coefficients (ADC) and T2w images (as shown in the first block of Fig. 1); 2) a dual-path multimodal CNN which takes the cropped and registered ADC-T2w image pair as input and produces a CS/nonCS classification score and a CS-class response map CRM_{CS} for each modality (as shown in the second block of Fig. 1). The pixel with a large response value on CRM_{CS} indicates a high probability of this pixel to be CS PCa. The CRM_{CS} of both modalities are then fused for accurate CS PCa localization (as shown in the third block of Fig. 1). During the training phase, all parameters of TDN and multimodal CNN are jointly optimized by minimizing three well-designed loss functions, i.e. CS vs. nonCS classification loss, inconsistency loss derived from the L2 distance between CRM_{CS} of the two modalities and overlap loss derived from the overlapping area of the two mutually excluded CRMs, i.e. CRM_{CS} and CRM_{nonCS} (as shown in the bottom part of Fig. 1). The three loss functions implicitly enforce high accuracy for prostate detection, ADC-T2w registration and extraction of representative PCa-relevant features. This is because mis-detection of prostate region, mis-alignment between ADC and T2w images and/or extraction of irrelevant PCa features could degrade classification accuracy, yield inconsistency between CRMs of the two modalities and lead to incorrect overlap between CRM_{CS} and CRM_{nonCS}.

II. Method

Diagnosing PCa using functional modalities (e.g. DWI, DCE-MRI) in addition to anatomical information, i.e. T2w, has been becoming an established clinical consensus [3] and widely used in many existing works [4], [6], [28]. Previous studies [4], [6] have demonstrated that combination of ADC computed from DWI and T2w can significantly increase both sensitivity and specificity of PCa detection. Thus, in this work we also use the ADC and T2w image modalities for the CS PCa detection task. Given pairs of ADC and T2w images, each of which is obtained from the same human body position, our goal is to localize CS PCa lesions whose GS ≥ 7, if any, in each image slice of mp-MRI data. Fig. 1 illustrates the framework of our system which consists of three modules: 1) the TDN for prostate detection and registration of ADC and T2w images based on TPS transformation [32] parameterized by regressed control points; 2) the dual-path multimodal CNN for CRM generation; and 3) the post-processing module which localizes CS PCa based on a final CRM_{CS} generated by combining the two CRM_{CS} from both modalities. In the following, we describe each module in details.

A. TDN for Prostate Detection and ADC-T2w Registration

To make the best use of ADC and T2w images for PCa detection, it is necessary to remove motion-induced misalignments and detect prostate region to exclude distracting surrounding tissues. In this section, we present the TDN for the prostate detection and registration tasks. The TDN can be embedded into any CNN-based PCa detection network and hence all parameters of TDN can be jointly optimized with the following PCa detection task in an end-to-end manner, yielding globally optimized parameters for all steps in the framework.

As shown in the first block of Fig. 1, the TDN consists of two parts: 1) two convolutional networks, each of which includes 2 convolutional layers (ConvLayer) and 2 fully connected layers (FCLayer), for regressing control points' locations on ADC and T2w images respectively, and 2) the thin plate spline transformation parameterized by the regressed control points for cropping and warping images of each modality. In the following, we detail each part.

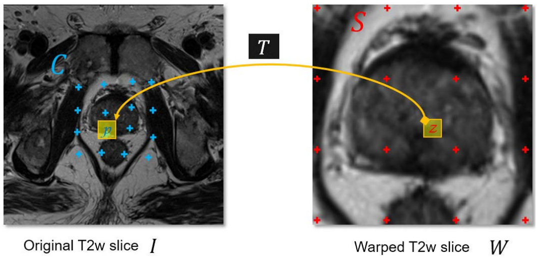

Fig. 2. A warped T2w slice *W* is derived from the original slice *I* using TPS transformation parameterized by *T* and bilinear sampling.

1) CNN for Regressing Control Points' Locations: Given an original slice I, TDN first resizes it to I' with a fixed size of 80×80, and then convolves I' with 8 convolutional kernels with size of 3×3 and 16 convolutional kernels with size of 3×3 in two ConvLayers sequentially. Both ConvLayers use 2 pixels stride and 1 pixel width zero padding, yielding a flattened 6400-dimensional feature. Before the final output, a FCLayer reduces the 6400-d feature to 1024-d intermediate representation, which is then fed into the last FCLayer, outputting x, y-coordinates of m control points. The regressed control points are called *target control points*, denoted by $\boldsymbol{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_m] \in \mathbb{R}^{2\times m}$. We normalize the value of coordinates to $[-1, 1]$ as $c_{i\_nx} = (2c_{i\_x} - w)/w$, $c_{i\_ny} = (2c_{i\_y} - h)/h$, where c_{i_nx} and c_{i_ny} are normalized x, y coordinates of a control point, c_{i_x} and c_{i_y} are unnormalized coordinates, w and h are width and height of the input image I. Outputs of each layer are all activated by the *tanh* function. In principal, the more control points are used, the greater degree of non-rigid transformation can be derived. In our experiments, we align a T2w image (i.e. the moving target) to the corresponding ADC image (i.e. the fixed reference) to avoid potential intensity artifacts on ADC images due to non-rigid image warping. Thus, we place more control points (i.e. $m = 16$) on T2w images for both cropping and warping and just enough control points (i.e. $m = 4$) on ADC images for only cropping prostate regions.

Note that the original image is grey-scale (single channel). However, to fit those images to the GoogLeNet, based on which our dual-path multimodal CNNs are built, which takes 3-channel inputs (i.e. color images), we adopted a regular routine which inputs the same grey-scale MR image into each of the RGB channels. Using this scheme, we can directly load weights from a pre-trained model with little modifications and thus minimize the non-trivial engineering burden of rebuilding a new architecture with little degradation to the final performance.

2) Image Cropping and Warping Based on TPS Transformation: Based on target control points, TPS transformation T is derived to crop the prostate region from an original slice and warp the cropped region accordingly, as shown in Fig. 2.

Specifically, we first define m *source control points*, denoted by $\boldsymbol{S} = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_m] \in \mathbb{R}^{2\times m}$, as indicated by red crosses in the right image of Fig. 2. Source control points are those at the intersections of regular, equal-spaced grids. As explained above, we use 4 source control points defined by a 2×2 grid for ADC images and 16 source control points defined by a 4×4 grid for T2w images. Based on pairs of corresponding source control points $\boldsymbol{S}$ and target control points $\boldsymbol{C}$, the TPS T is computed as:

$$\boldsymbol{T} = \left(\boldsymbol{L}^{-1} \cdot \begin{bmatrix} \boldsymbol{C}^T \\ \boldsymbol{0}^{3\times 2} \end{bmatrix}\right)^T \in \mathbb{R}^{2\times\ (m+3)} \qquad (1)$$

where $\boldsymbol{L}$ is derived from source control points as:

$$\boldsymbol{L} = \begin{bmatrix} \tilde{\boldsymbol{S}}^T & \boldsymbol{R} \\ \boldsymbol{0}^{3\times 3} & \tilde{\boldsymbol{S}} \end{bmatrix} \in \mathbb{R}^{(m+3)\times\ (m+3)} \tag{2}$$

where $\tilde{\boldsymbol{S}} = \begin{bmatrix} \boldsymbol{1}^{m\times 1} & \boldsymbol{S}^T \end{bmatrix}^T \in \mathbb{R}^{3\times m}$ is the homogeneous coordinates of $\boldsymbol{S}$ and $\boldsymbol{R}$ is a symmetric matrix, each entry is $r_{i,j} = \phi(\left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|_2)$, where $\phi(d) = d^2\ ln(d^2)$ is the radial basis function (RBF). Based on the estimated TPS transformation each pixel z in a warped image W can find its corresponding point p in the original slice I as shown in Fig. 2. Specifically, for each pixel $\tilde{z} = \begin{bmatrix} 1, & z^T \end{bmatrix}^T$ in W, where $z = [x, y]^T$ is the x, y-coordinates of the pixel, the corresponding point $\boldsymbol{p}$ in I is:

$$\boldsymbol{p} = \boldsymbol{T} \cdot \begin{bmatrix} \tilde{z}^T & \phi(\|z - \boldsymbol{s}_1\|_2) & \cdots & \phi(\|z - \boldsymbol{s}_m\|_2) \end{bmatrix}^T \in \mathbb{R}^2 \tag{3}$$

Applying Eq. 3 on every single pixel in W, we can obtain a set of corresponding points $\boldsymbol{P} = \{\boldsymbol{p}_i | i = 1, 2, \ldots, N\}$ in I, where N is the total number of pixels in W. Therefore, the value of each pixel z in W can be constructed from the values of pixels around the corresponding point $\boldsymbol{p}$ in I by a bilinear sampler as $W = BiliSamp\ (\boldsymbol{P}, I)$. Theoretically, the desirable image W can have an arbitrary size after bilinear sampling, we constrain W to have a fixed size of 299 × 299 to satisfy the requirement of input size of the subsequent PCa detection CNN. Both TPS transformation and bilinear sampling have been demonstrated in [33] to be differentiable so that errors can be back-propagated for target control points regression in the TDN.

It is worth noting that the control points on individual modalities are only used as the parameters of TPS for cropping and warping and have no semantic meaning of anatomical landmarks. Additionally, an individual TPS transformation is calculated for images of a single modality and the number of control points among all images of the same modality is consistent.

3) TDN Optimization: As denoted by green arrows in Fig. 1, the optimization of the TDN is embedded into the training procedure of the subsequent dual-path multimodal CNN by minimizing three types of loss functions, i.e. classification loss, inconsistency loss and overlap loss. More details of the loss functions will be provided in Sec. II-B). The smaller mis-alignment errors between ADC and T2w images and more accurate detection of prostate regions in ADC and T2w images, the easier for the multimodal CNN to achieve smaller values for the three loss functions. Therefore, during the training phase the locations of $\boldsymbol{C}$ for ADC and T2w images are automatically adjusted according to values of the three loss functions back-propagated from the subsequent dual-path multimodal CNN. In summary, there is a strong trend for the TDN to automatically regress two proper sets of control points in ADC and T2w images to minimize their mis-alignment errors and detect proper prostate regions even in the absence of supervised information, e.g. anatomical landmarks. It is worth mentioning that [33] and [34] adopted a similar idea that enables the traditional CNNs to be of spatial attention by using the TPS transformation. However, both works used TPS for improving the classification/recognition accuracy of single-modality images, while our method is the first try for multimodal registration and for facilitating an end-to-end system for CS PCa detection.

To facilitate a faster regression speed, in practical implementation we also provide supervised annotations based on bounding boxes roughly denoting prostate regions which are relatively easy to obtain. Specifically, we define $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4] \in \mathbb{R}^{2\times 4}$ the x, y-coordinates of the four corners of the bounding box roughly indicating a prostate region, the regressed four control points (i.e. four outmost corners of the grids) $\boldsymbol{C}$ can be derived by minimizing the regression loss as:

$$\ell_{reg,ADC}(\boldsymbol{C}_{ADC}, \boldsymbol{Y}_{ADC}) = \frac{1}{4}\sum_{i}^{4} \|\boldsymbol{c}_i - \boldsymbol{y}_i\|_2^2 \tag{4}$$

$$\ell_{reg,T2w}(\boldsymbol{C}'_{T2w}, \boldsymbol{Y}_{T2w}) = \frac{1}{4}\sum_{i}^{4} \|\boldsymbol{c}'_i - \boldsymbol{y}_i\|_2^2 \tag{5}$$

Roughly annotating bounding boxes encompassing prostate regions require lower level of expertise. This procedure can even be automated by roughly placing a 80 × 80 square in the middle of an 256 × 256 image slice according to the prior knowledge of the relative location and size of a prostate in abdominal scans.

B. Dual-Path Multimodal CNN for CRM Generation

In this section, we present a dual-path multimodal CNN trained in a weakly-supervised manner for class response map generation. The dual-path multimodal CNN is built based on our previous work [35] which consists of two parallel single-modality CNNs, each of which is trained based on only image-level annotations (i.e. presence/absence of CS PCa), yet can learn effective CS PCa-relevant features from cluttered prostate and surrounding tissues. In the following, we first briefly describe the original design followed by our modifications.

1) Original Dual-Path Multimodal CNN: The original design consists of two single-modality CNNs for ADC and T2w image respectively. Each single-modality CNN is a truncated GoogLeNet composed of a batch of Inception V1 modules [36]. The outputs of the truncated GoogLNet are 1024 feature maps of size 16 × 16 each. We convolve the 1024 feature maps with a kernel size of 1 × 1 × 1024 to obtain a single feature map to which we apply the global average pooling (GAP) operation to produce a single score output. We project the output score to the range [0, 1] by the Softmax function, which is then used as the final output indicating the probability of the input image to be cancerous. To combine the information from both ADC and T2w images, a new loss function, called inconsistency loss, is designed in [35] which is the L2 distance between the final feature maps of ADC (i.e. M_{ADC}) and T2w (i.e. M_{T2w}) calculated as follows:

$$\ell_{inc}(M_{ADC}, M_{T2w}) = \frac{1}{N}\|\sigma(M_{ADC}) - \sigma(M_{T2w})\|^2 \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, N is total number of pixels in the feature map, and $\|\sigma(M_{ADC}) - \sigma(M_{T2w})\|$ calculates

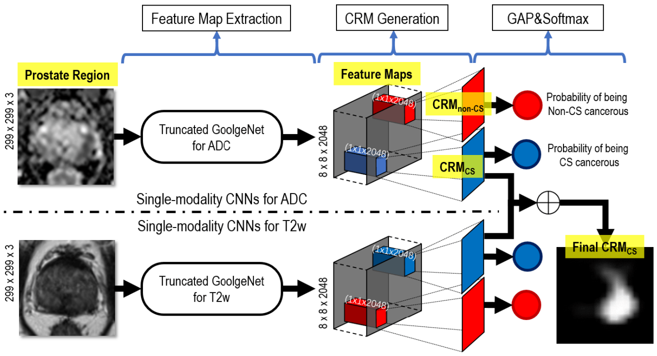

Fig. 3. Architecture of our dual-path multimodal CNN for CS vs. nonCS classification and CRM generation.

the pixel-wise differences between the M_{ADC} and M_{T2w}. The training process of the original dual-path multimodal CNN is to optimize both classification accuracy according to the ground truth image-level labels for both ADC and T2w images and the inconsistency loss between M_{ADC} and M_{T2w}. That is, the training minimizes the following back-propagate error E:

$$E = w_1 \left(\ell_{cls}(p_{ADC}, y) + \ell_{cls}(p_{T2w}, y)\right) + w_2\, \ell_{inc}(M_{ADC}, M_{T2w}) \quad (7)$$

where $\ell_{cls}(p_{ADC}, y)$ and $\ell_{cls}(p_{T2w}, y)$ are classification losses of ADC and T2w respectively, defined as:

$$\ell_{cls}(p, y) = -[\ y\ log(p) + (1 - y)\ log(1 - p)\ ] \quad (8)$$

where p is output class probability and y is the class label; w_1 and w_2 are weights for adjusting the contributions of the classification loss and inconsistency loss to E for an optimized CNN model.

2) Modifications: The enhanced dual-path multimodal CNN is shown in Fig. 3. Three important modifications have been made for further improvement:

- For each single modality we adopt a truncated GoogLeNet composed of a batch of Inception V3 modules [37] rather than Inception V1 modules for its better classification accuracy and efficiency. The truncated GoogLeNet outputs 2048 feature maps with a size of 8 × 8 each.
- Instead of generating a single feature map for each modality in our original design, in this work we convolve the 2048 feature maps using two separate convolutional kernels size of 1 × 1 × 2048 to generate two feature maps. The two feature maps can be considered as the class response maps (CRM) for CS PCa class (CRM$_{CS}$) and nonCS PCa class (CRM$_{nonCS}$), with each pixel on the map indicating the likelihood of this pixel belonging to the corresponding class. We apply GAP to both CRM$_{CS}$ and CRM$_{nonCS}$ to generate two scores (s_{CS} and s_{nonCS}) for the two classes. s_{CS} and s_{nonCS} are then normalized to [0, 1] using the Softmax function as $p_{CS} = e^{s_{CS}} / \left(e^{s_{CS}} + e^{s_{nonCS}}\right)$ and $p_{nonCS} = e^{s_{nonCS}} / \left(e^{s_{CS}} + e^{s_{nonCS}}\right)$ to denote the probabilities of the input image belonging to the two classes (i.e. CS or nonCS) respectively.
- We design a new loss function, i.e. overlap loss, which is calculated by computing the area of the overlapped regions between CRM$_{CS}$ and CRM$_{nonCS}$:

$$\ell_{overlap} = \frac{1}{N} \sum_{x,y} (\text{CRM}_{CS} \cap \text{CRM}_{nonCS}) \quad (9)$$

where N is the total number of entries in a CRM. The value at location (x, y) for CRM$_{CS} \cap$ CRM$_{nonCS}$ is defined as $min(\text{CRM}_{CS}(x, y), \text{CRM}_{nonCS}(x, y))$. The CRM$_{CS} \cap$ CRM$_{nonCS}$ is a map approximately representing the overlapped regions between CRM$_{CS}$ and CRM$_{nonCS}$. The new loss is designed based on the fact that CS cancerous tissues should have little overlapped regions with nonCS tissues, and thus the overlapped regions between the CRM$_{CS}$ and CRM$_{nonCS}$ of an input should be zero. We integrate the overlap loss into the back-propagate error E, and by minimizing the overlap loss in addition to the other loss functions, the network can be trained to better capture CS PCa-relevant features and suppress irrelevant features.

3) Dual-Path Multimodal CNN Optimization: We optimize the dual-path CNN in a weakly-supervised manner by minimizing a weighted summation of three types of losses defined above:

$$E = w_1 \left(\ell_{cls,ADC} + \ell_{cls,T2w}\right) + w_2 \left(\ell_{inc,CS} + \ell_{inc,nonCS}\right) + w_3 \left(\ell_{overlap,ADC} + \ell_{overlap,T2w}\right) \quad (10)$$

where w_1, w_2 and w_3 are manually tuned weights to scale the three types of losses to the same scale meanwhile adjust the contributions of losses to the training. In our experiments we set them as $w_1 = 1$, and $w_2 = w_3 = 0.2$. During the training phase, only image-level annotations indicating the presence/absence of PCa are demanded. The reason why minimizing the back-propagate error E defined in Eq. 10 can effectively enable proper prostate detection, registration and accurate CS PCa detection is as follows.

Based on prior studies we know that entries in a CRM correspond to small patches in an input image to which they are path-connected. Such small patches are also called *receptive fields* [38]. Entries of CRM$_{CS}$ will respond strongly (i.e. have a large response value) if their receptive fields contain visual patterns which are discriminative and consistent across all training data for contributing to the classifcation of the slice as CS cancerous, and so is CRM$_{nonCS}$. During the back propagation (BP) procedure, in order to minimize the classification losses and overlap losses, parameters of the CNN will be updated to diminish responses of entries in CRM$_{CS}$ for nonCS slices and increase responses of entries in CRM$_{CS}$ whose receptive fields contain CS cancerous lesions for CS slices. Similar explanation is also applicable to CRM$_{nonCS}$. Therefore, by providing only image-level annotations we can still expect reasonable CRM$_{CS}$/CRM$_{nonCS}$, each entry of which is essentially the likelihood of this position belonging to the corresponding class (i.e. CS/nonCS). The effectiveness of above described weakly-supervised learning has been demonstrated in recent works [35], [39], [40].

Since the CS PCa-relevant visual patterns appear at the same locations in registered ADC and T2w images while the irrelevant visual patterns could scatter over different places. Minimizing inconsistency loss which explicitly enforces spatial consistency between CRMs of ADC and T2w can effectively guide the CNN learning process to suppress irrelevant visual patterns and highlight CS PCa-relevant patterns which are consistent in both ADC and T2w CRMs. Additionally, minimizing the inconsistency loss is also beneficial to image registration based on the TDN. This is because inconsistency between CRMs generated by the two independent single-modal CNNs is partially caused by misalignments between unregistered ADC-T2w slices. Minimizing the inconsistency loss can guide the learning process of the TDN towards regression of more accurate matching of target control points for registration.

C. Post-Processing for Accurate CS PCa Localization

In testing, when an ADC-T2w slice pair is classified as positive, we localize CS PCa on the corresponding CRM$_{CS}$ via three-step post-processing. First, to avoid performance degradation arising from the low resolution of CRM$_{CS}$ (i.e. 8 × 8) we adopt a well-known solution called 'shifting&stitching' [38], [41] to up-sample the CRM$_{CS}$ to the same size as the input of the multimodal CNNs (i.e. 299 × 299). Specifically, we generate 100 input images for the multimodal CNNs, each of which is obtained by shifting the obtained image by the TDNs (by left and top padding) x pixels to the right and y pixels down, where $(x; y) \in \{0, 1, \ldots, 9\} \times \{0, 1, \ldots, 9\}$. Each of the sifted inputs runs through the multimodal CNN, and the 100 outputs are interlaced to form a CRM with a size of 80 × 80, which is then resized to 299 × 299 for CS PCa localization.

Second, we fuse CRM$_{CS,ADC}$ and CRM$_{CS,T2w}$ to generate a single CRM$_{CS}$ by pixel-wise summation. We believe the fused CRM$_{CS}$ can facilitate a more accurate CS PCa localization since ADC and T2w images can provide complementary information for CS PCa localization and adding CRM$_{CS,ADC}$ and CRM$_{CS,T2w}$ can further highlight regions with great response values in both ADC and T2w images. We experimentally demonstrate that the fused CRM$_{CS}$ can provide superior localization accuracy than CRM$_{CS,ADC}$ and CRM$_{CS,T2w}$. Specifically, the sensitivity achieved based on CRM$_{CS,T2w}$, CRM$_{CS,ADC}$ and the fused CRM$_{CS}$ are 86.6%, 90.7% and 96.9% respectively, and the number of false alarms are 14, 12 and 5 respectively.

Third, we find local maximums in the fused CRM$_{CS}$ whose response values are larger than its neighboring entries as suspicious CS PCa pixels. We further filter out outliers from all suspicious CS PCa pixels via an adaptive thresholding method based on the Otsu's [42]. Specifically, a threshold is automatically selected based on the Otsu's algorithm for an input fused CRM$_{CS}$. Suspicious CS PCa pixels whose responses are smaller than the threshold are excluded as outliers.

III. Data Preparation and Evaluation Metrics

A. Patient Characteristics and Image Acquisition Protocol

The study was approved by our local institutional review board. The mp-MRI data used in the study are collected from two datasets: i) a locally collected dataset named TJPCa Dataset which includes data of 156 patients conforming to the following five criteria: 1) the data for PCa assessment were acquired between June 2014 and December 2015; 2) all the data include either pathologically-proven PCa or benign prostatic hyperplasia (BPH) by a 12-core systematic TRUS-guided plus targeted prostate biopsy which were performed within six weeks after the MRI examination; 3) the data is from the patients who did not received focal therapy, hormones, or radiation prior the MRI scan, 4) the data include both ADC and T2w images, and 5) the imaging data does not include severe artifacts that made the examination non-diagnostic. Indications for prostate MRI include: tumor detection for patients with clinical suspicion of prostate cancer (elevated PSA > 4.0 ng/mL and/or suspicious DRE) before biopsy, cancer staging, radiation planning, surgical planning, active surveillance, planning for biopsy targeting and evaluation of patients with a prior negative biopsy but could have continuous clinical suspicion of prostate cancer. ii) a publicly available dataset named the PROSTATEx (training), which is the training set from the PROSTATEx challenge [3], [43], [44], including data of 70 MRI-targeted biopsy-proven CS PCa and 134 nonCS PCa patients. Remaining testing data of 140 patients from the PROSTATEx challenge are excluded from this study due to the lack of ground-truth labels (i.e. CS or nonCS). Table I provides detailed characteristics of the two datasets.

TABLE I
Patients Characteristics From TJPCa Dataset and PROSTATEx (Training) Dataset

Type	Characteristics	TJPCa	PROSTATEx (training)
CS PCa	Total number of patients	64	70
	Mean age (years old)	66.6 ± 8.5	–
	Age range (years old)	50 — 88	–
	Median PSA value (ng/ml)	53.8	–
	PSA values range (ng/ml)	4.6 — 1,000	–
	Biopsy proven	✓	✓
	Gleason score	7 — 10	–
nonCS PCa	Total number of patients	92	134
	Mean age (years old)	69.0 ± 8.4	–
	Age range (years old)	51 — 85	–
	Median PSA value (ng/ml)	11.8	–
	PSA values range (ng/ml)	0.26 — 168.8	–
	Biopsy proven	✓	✓

All mp-MRI images in the TJPCa dataset were acquired between June 2014 and December 2015 on a 3.0 Tesla (T) whole-body unit MR imaging system (MAGNETOM Skyra, Siemens Medical Solutions, Erlangen, Germany), running software version Syngo MR D13. The acquisition parameters for the transverse, coronal, and sagittal T2WI TSE images were set as follows: repetition time [TR] is 6750 ms, echo time [TE] is 104 ms, echo train length is 16, section thickness is 3 mm, there is no intersection gap, field of view [FOV] is 180 mm and the image size is 384 × 384. The acquisition parameters for the transverse plane of DWI sequences were set as follows: b values are 0 and $1,000$ s/mm^2, TR/TE are 6750 ms/ 104 ms, section thickness is 3 mm, FOV is 180 mm and the image size is 180 × 144. The ADC maps were computed from an Advanced Workstation. MRI protocols for data acquisition of the PROSTATEx (training) dataset are provided in [3].

B. Data Preparation and Evaluation Metrics for 5-Fold Cross-Validation

1) Data Preparation: We evaluate the performances of image-level CS vs. nonCS classification and CS lesion localization based on 5-fold cross-validation (CV). For doing so, we evenly divide the data of a total of 360 patients (134 CS PCa patients and 226 nonCS PCa patients) into 5 parts in terms of patient and each part contains around 27 CS PCa patients and 45 nonCS PCa patients. From each CS PCa patient data, 3 to 5 ADC-T2w pairs were manually selected by the radiologist where CS lesions are clearly visible and the cross-section area of each lesion is sufficiently large. From each nonCS PCa patient data, 10 to 13 ADC-T2w pairs were selected where prostate gland is clearly visible. As a result, we obtained a total of around 600 original ADC-T2w pairs (100 CS and 500 nonCS ADC-T2w pairs) in each part for 5-fold CV. In each round of 5-fold CV, 4 parts are used for training and the remaining one is used for testing. To prevent the impact of imbalance data and provide sufficient images for training, we apply similar data augmentation strategy as [35] to augment positive pairs (i.e. image pair contains PCa with GS ≥ 7) by 20 times and negative pairs (i.e. image pair is normal, BPH or PCa with GS ≤ 6) by 4 times for the training. Every image pair has both image-level label indicating presence/absence of CS lesions and pixel-level label indicating the specific regions of CS lesions in T2w image by two experts' manual annotations.

To provide a proper evaluation of the registration results, we seek assistance from a radiologist to manually label 145 prostate gland regions in ADC-T2w image pairs. To delineate the boundary of a prostate gland in an ADC map, the radiologist first labeled the prostate gland in a T2w image, and the delineation on the T2w image is mapped to the corresponding ADC map as a reference. Then, a radiologist carefully adjusted the location, angle and size of the delineation on the ADC map to ensure that it tightly encloses the prostate gland in the ADC map.

2) Evaluation Metrics for CS Localization: We used free-response receiver operating characteristic (FROC) curve [45] as [3] for evaluating the performance of CS PCa localization. In our case, each point in the FROC curve indicates a lesion localization fraction (LLF) vs. non-lesion localization fraction (NLF) at a given threshold, where LLF is defined as the number of successfully localized CS lesions (i.e. true CS lesions with localized points in their reference standard annotation) divided by the total number of true CS lesions, and NLF is defined as the total number of localized points falling on nonCS tissues divided by the total number of BPH/normal patients. We expect a better CS localization system to achieve a higher LLF (i.e. sensitivity) at a given NLF (i.e. false positives per BPH/normal patient). Therefore we also used the sensitivity at 0.1 NLF (denoted by *Sensi@0.1NLF*) and the sensitivity at 1.0 NLF (denoted by *Sensi@1.0NLF*) to evaluate the performance of CS localization.

3) Evaluation Metrics for CS vs. nonCS Classification: Since in clinical practice the first task is to determine whether a patient has CS PCa or not while successful CS localization is the second step, we also analyze the performance of image-level CS vs. nonCS classification using the receiver operating characteristic (ROC) curve. In addition to the ROC curve, we evaluate the classification performance using a common metric, namely area under curve (AUC). Since the evaluation is based on 5-fold CV, for each metric, we report the average of the results of five folders. We also report the standard deviations of AUC, Sensi@0.1NLF and Sensi@1.0NLF. Furthermore, statistical significance testing based on the permutation testing, as suggested by [46], is performed to evaluate the statistical significance of the improvements achieved by our method.

4) Evaluation Metrics for Multimodal Registration: we evaluated the registration performance using two widely-used metrics in previous studies [47], [48]: 1) Dice Similarity Coefficient (DICE) and 2) Hausdorff distance (HD). Specifically, DICE is calculated as:

$$\frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{11}$$

where X and Y are prostate masks of the fixed (i.e. original ADC manual segmentation mask of a prostate) and registered images (i.e. registered T2w manual segmentation mask of the prostate). HD is calculated as:

$$\max\left\{\max_{u \in L_{ADC}} \min_{v \in L_{T2w}} d(u, v), \max_{u \in L_{T2w}} \min_{v \in L_{ADC}} d(u, v)\right\} \tag{12}$$

where L denotes a binary mask of a prostate. $d(u, v)$ is the Euclidean distance between pixels u and v. An accurate registration result should have a high DICE and a low HD.

IV. Results

A. Evaluation of Multimodal Registration

We first compare the performance of multimodal registration among three approaches: *TDN-d*, *MI-r* + *TDN-d* and *TDN-r*. The TDN-d is trained by regressing only 4 control points for both ADC and T2w images, which can thus only detect prostate without the ability for ADC-T2w image registration. The MI-r + TDN-d first aligns a T2w image to its corresponding ADC image by a MI-based multimodal registration method [49] using affine transformation, and then detects the prostate in the registered T2w image by the TDN-d, at last the detected bounding box of prostate is projected to the ADC image. The TDN-r is the complete version of our proposed framework, regressing 4 control points on ADC images and 16 control points on T2w images, which can therefore achieve both prostate detection and registration. We also report the DICE and HD values for the data without prostate detection and multimodal registration (denoted by *Raw Data*).

For a fair comparison, we optimize the mutual information based registration method (i.e. MI-r) by exhaustively searching for the optimized parameters which could provide the highest average DICE value on the 145 ADC-T2w pairs. Table II shows the comparison results.

TDN-r outperforms both TDN-d and MI-r + TDN-d, achieving a greater average DICE and a lower HD. Moreover, the MI-based method requires an exhaustive search of proper parameters over the dataset for optimized registration results.

TABLE II
THE COMPARISON RESULTS (AVG ± STD DEV) AMONG THREE DIFFERENT APPROACHES

Methods	DICE %	HD(mm)
Raw Data	79.8±2.7	11.5±3.3
TDN-d	88.5±2.1	8.5±2.8
MI-r+TDN-d	92.8±2.2	7.2±2.3
TDN-r	**94.2±1.7**	**6.7±1.8**

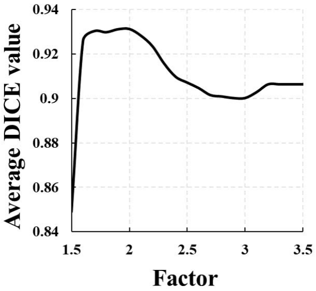

Fig. 4. AverageDICE value by MI-r + TDN-d with respect to a tuning factor which is the search radius divided by a default radius (i.e. 6.25×10^{-3}).

Fig. 4 shows the average DICE number with respect to a tunable parameter for the MI-r: the initial search radius used in the optimizer of MI-r. An improper radius size could lead to over 3% performance drop. Even though the parameter optimization process can be automated, in practice the parameters optimized for one dataset might not achieve satisfactory performance on images with different characteristics (e.g. acquired using different scanners/protocols). In contrast, our proposed TDN-r does not require any parameter tuning for optimized registration, yielding greater convenience and broader applicability for clinical usage.

B. Demonstration of the TDN-r

In addition to the demonstration of the TDN-r for accurate multimodal registration, we analyze the impact of the TDN-r on the overall performance of our CAD system. Since the TDN can successfully detect all prostates in testing, we present only comparison results of image-level classification and CS lesion localization among three models: *TDN-r*, *TDN-d* and *MI-r + TDN-d*. To obtain the final classification and localization results, we concatenate each model with the proposed dual-path multimodal CNN and train it by minimizing losses including classification loss, inconsistency loss and overlap loss.

As observed from Figs. 5(a) and (b), all three models have better performance of image-level classification on ADC images than T2w images. Therefore, we only provide results on ADC images when comparing performance of image-level CS vs. nonCS classification for the following evaluations. Table III shows the comparison results among the three models of both image-level classification and CS lesion localization tasks. For image-level CS vs. nonCS classification, our multimodal registration based on TDN-r achieves higher AUC ($p < 0.007$) than those based on MI-r. Compared with TDN-d, the AUC achieved by our method is 3.5% higher ($p < 0.005$). For CS lesion localization, our method is better than MI-r, achieving 8.3% higher at Sensi@0.1NLF ($p < 0.02$) and 7.2% higher at Sensi@1NLF ($p < 0.02$), and significantly better than TDN-d, achieving 40% higher at Sensi@0.1NLF ($p < 0.02$) and 14% higher at Sensi@1NLF ($p < 0.001$). These results demonstrate that the complete version of the TDN, i.e. TDN-r, can well align T2w to ADC images and help achieve significantly better results for classification and localization than both TDN-d and MI-based method.

C. Demonstration of Multimodal Fusion

In this section, we exam the performance of our multimodal fusion method and compare the results with those achieved by single-modal CNNs and another two commonly used fusion strategies.

1) Single-modal CNN vs. Multimodal CNN: We compare our multimodal fusion method with the two single-model CNNs for ADC (denoted as *ADC Single*) and T2w (denoted as *T2w Single*) which are trained independently based on images of a single modality. Although there is no need for registration, we use TDN-r for prostate detection in each single modality for a fair comparison.

As shown in Figs. 6(a) — (b) and the 2^{nd} column in Table IV, for image-level CS vs. nonCS classification, our proposed method with multimodal fusion achieves better performance than those with single modality ($p < 0.01$ for ADC Single, $p < 0.001$ for T2w Single). As shown in Figs. 6(c) and the 3^{rd} — 4^{th} columns in Table IV, for the CS lesion localization, our proposed method with multimodal fusion significantly outperformed those with single modality by 10% to 47% as shown in the fourth and fifth columns of Table IV (Sensi@0.1NLF and @1NLF: $p < 0.001$). These results demonstrate that fusing ADC and T2w information based on our method can effectively highlight true CS lesions and suppress nonCS tissues. For example, when only one false positive is allowed per 10 normal patients, our system can successfully localize 63.7% CS lesions, while both *ADC Single* and *T2w Single* could miss more than 70% CS lesions, which is definitely unacceptable in clinical practice.

2) Comparison With Other Strategies of Multimodal Fusion: We compare our fusion strategy with the two widely-used multimodal fusion strategies: i) directly concatenating two 2048 feature maps into 4096 feature maps, which are then convolved by two convolutional kernels with a size of $1 \times 1 \times 4096$ to generate the final CRM_{nonCS} and CRM_{CS}, as illustrated in Fig. 7(b), and ii) pixel-wise summation of two 2048 feature maps to form 2048 feature maps, which are then convolved by two convolutional kernels with size of $1\times1\times2048$ to generate the final CRM_{nonCS} and CRM_{CS}, as illustrated in Fig. 7(c). In training, models of the two multimodal fusion strategies are trained to minimize the classification losses and the overlap loss. The predictions are obtained by applying GAP and the Softmax function to their CRMs.

Table V and Fig. 8 show the comparison results. The results clearly indicate that our method is better than other multimodal fusion methods for image-level CS vs. nonCS classification ($p < 0.005$ for both Concatenate and Sum).

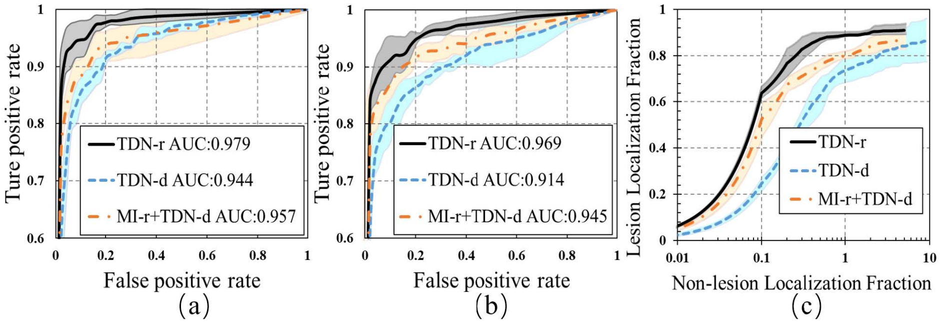

Fig. 5. Comparison results based on 5-fold cross-validation. 95% confidence intervals are shown as transparent areas around the mean curves. (a) and (b) Average ROC curves of CS vs. nonCS classification applying on ADC and T2w images respectively and (c) average FROC curves of CS lesion localization based on the three models: TDN-r, TDN-d and MI-r + TDN-d.

TABLE III

COMPARISON RESULTS (AVG ± STD DEV) AMONG THREE MULTIMODAL REGISTRATION METHODS BASED ON 5-FOLD CROSS-VALIDATION

	CS vs. nonCS Classification	CS Lesion Localization	
	AUC	Sensi@0.1NLF	Sensi@1.0NLF
TDN-d	0.944 ± 0.010	0.236 ± 0.038	0.757 ± 0.064
MI-r+TDN-d	0.957 ± 0.011	0.554 ± 0.085	0.826 ± 0.059
TDN-r	**0.979** ± **0.009**	**0.637** ± **0.023**	**0.898** ± **0.021**

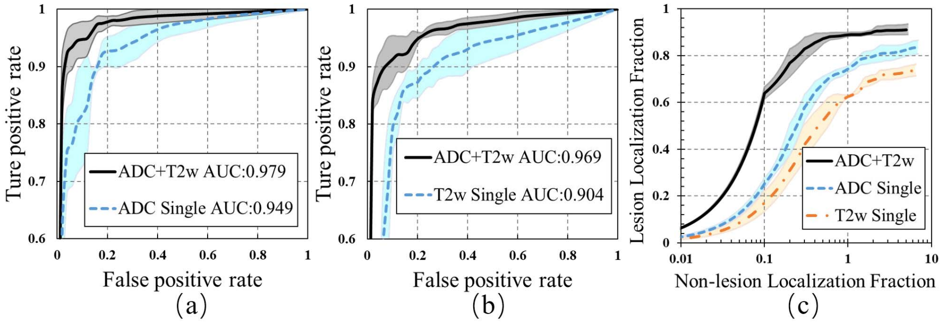

Fig. 6. Comparison results based on 5-fold cross-validation. 95% confidence intervals are shown as transparent areas around the mean curves. (a) and (b) Average ROC of CS vs. nonCS classification for ADC and T2w images respectively based on single-modal CNNs for ADC and T2w and our multimodal CNN. (c) Average FROC of CS PCa localization based on single-modal CNNs for ADC and T2w and our multimodal CNN.

TABLE IV

COMPARISON RESULTS (AVG ± STD DEV) AMONG SINGLE-MODAL CNNS AND OUR MULTIMODAL CNNS BASED ON 5-FOLD CROSS-VALIDATION

	CS vs. nonCS Classification	CS Lesion Localization	
	AUC	Sensi@0.1NLF	Sensi@1.0NLF
ADC Single	0.949 ± 0.010	0.240 ± 0.036	0.764 ± 0.047
T2w Single	0.904 ± 0.012	0.161 ± 0.044	0.617 ± 0.018
ADC+T2w	**0.979** ± **0.009**	**0.637** ± **0.023**	**0.898** ± **0.021**

TABLE V

COMPARISON RESULTS (AVG ± STD DEV) AMONG THREE DIFFERENT MULTIMODAL FUSION STRATEGIES BASED ON 5-FOLD CROSS-VALIDATION

	CS vs. nonCS Classification	CS Lesion Localization	
	AUC	Sensi@0.1NLF	Sensi@1.0NLF
Sum	0.955 ± 0.014	0.088 ± 0.029	0.601 ± 0.047
Concatenate	0.942 ± 0.012	0.204 ± 0.085	0.750 ± 0.042
Ours	**0.979** ± **0.009**	**0.637** ± **0.023**	**0.898** ± **0.021**

For CS lesion localization, our proposed method also achieves significantly better results than other methods (Sensi@0.1NLF and @1NLF: $p < 0.0005$ for both Concatenate and Sum). Though Sum slightly outperforms Concatenate, it has the worst performance for CS PCa localization which is even worse than *T2w Single*'s. We think the reason of poor performance of pixel-wise summation is that the summation of both CS PCa-relevant and -irrelevant features helps yield a

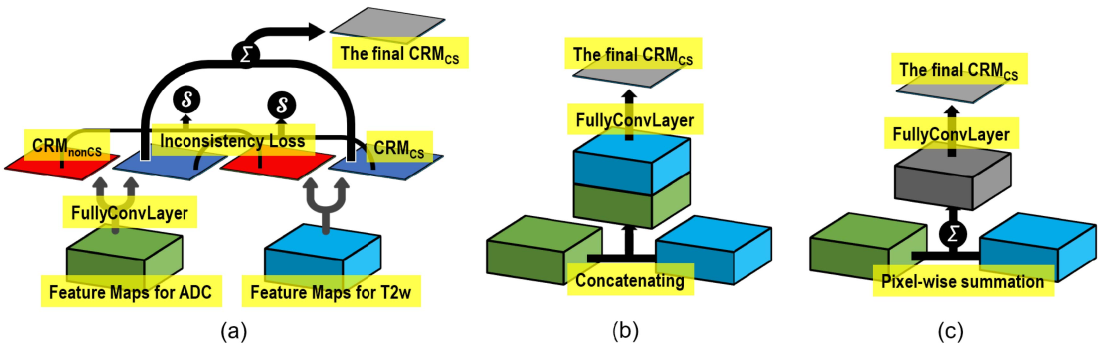

Fig. 7. Illustrations of three different multimodal fusion strategies: (a) ours based on the inconsistency loss to force the spatial consistency between CRMs of ADC and T2w, (b) directly concatenating two feature maps of ADC and T2w, and (c) pixel-wise summation of two feature maps of ADC and T2w.

better classification result, but those irrelevant visual patterns significantly degrade the performance of CS localization. The localization performance of Concatenate is quite similar to that of *ADC Single* whose results are shown in the first row of Table IV. The explanation is that with concatenation, the resulting CRM is a weighted sum of feature maps from ADC and T2w, and the CNN does not actually ‘fuse’ features from the two modalities. But it inclines to use ADC features for the final prediction by assigning greater weights to them, and in turn yields similar performance to that of the single-modal CNN for ADC. To summarize, our multimodal fusion scheme can effectively ‘fuse’ the two modalities by guiding the learning process of each modality to mutually affecting each other, which effectively removes unwanted visual patterns and captures the true CS lesions, and in turn achieves the best performance of CS localization.

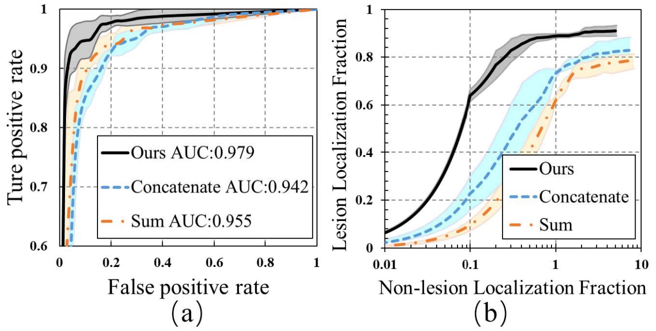

Fig. 8. Evaluation results among three multimodal fusion strategies based on 5-folder cross-validation. 95% confidence intervals are shown as transparent areas around the mean curves. (a) Average ROC curves of image-level CS vs. nonCS classification. (b) Average FROC curves of CS PCa localization.

D. Demonstration of Joint Optimization

We compare our system of CS PCa detection with the two baselines built based on state-of-the-art deep learning based methods, 2D U-Net [50] and 3D DeepMedic [51] which are proposed for automated lesion detection/segmentation. Specifically, DeepMedic analyzes 3D mp-MRI data and predicts a cancer response map indicating the probability of each voxel being CS PCa. DeepMedic fuses two modalities (i.e. ADC and T2w) by concatenating 3D input data of two modalities in the input channel, and employs a dense-training strategy for training. Similarly, the 2D U-Net processes a 2D image and outputs the probability of being cancerous for each image pixel. However, U-Net was originally designed to process images of a single modality. For a fair comparison, we applied the identical multimodal fusion method used in our framework, i.e. the inconsistency loss, to fuse ADC and T2w information in U-Net. In particular, two U-Nets for ADC and T2w respectively are trained concurrently by minimizing both segmentation error loss (original loss of U-Net) and inconsistency loss (for multimodal fusion).

We used the same set of 2D training images for training our model and the U-Net model. As DeepMedic requires 3D input, we used the 3D data, from which our 2D training images were selected, for training the DeepMedic model. When testing in each round of 5-fold CV, we first used MI-r + TDN-d to align and detect prostates in T2w-ADC image pairs and then U-Net/DeepMedic to generate CRM$_{CS}$. Once the 2D slice is identified as positive by our proposed dual-path multimodal CNNs, its corresponding CRM$_{CS}$ is used for CS lesion localization by the post-processing. Steps including MI-r + TDN-d, U-Net/DeepMedic, dual-path multimodal registration CNNs in the baseline methods are independent and optimized separately. Table VI and Fig. 9 show the comparison results.

The results indicate that our method with joint optimization outperforms the 2D U-Net by 41.2% at Sensi@0.1NLF ($p < 0.0005$) and by 6.5% at Sensi@1.0NLF ($p < 0.02$) respectively. Compared to 3D DeepMedic, our results is 38% better at Sensi@0.1NLF ($p < 0.0005$) and 12.5% better at Sensi@1.0NLF ($p < 0.0005$). The achieved superior performance of CS lesion localization by our proposed method demonstrates that joint optimization in our system makes learning processes of steps mutually benefit each other, leading to more accurate CS lesion detection than the individual optimized method. Fig. 10 presents 4 samples of localization results, where the baselines localize more false positives than ours.

E. Performance on the PROSTATEx Challenge

The PROSTATEx challenge focuses on quantitative image analysis methods for the classification of CS PCa from nonCS

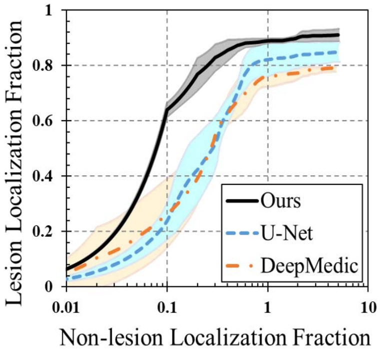

Fig. 9. Average FROC curves of CS PCa localization for two baseline methods, 2D U-Net and 3D DeepMedic, and our method based on 5-fold cross-validation. 95% confidence intervals are shown as transparent areas around the mean curves.

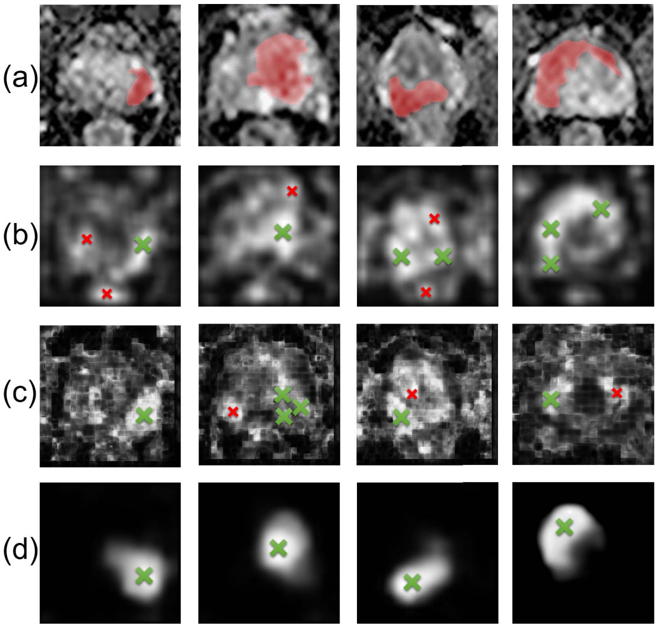

Fig. 10. (a) Prostate regions in ADC overlaid with manual delineation of CS lesion. (b)—(d) show CRM_{CS} generated by the U-Net, DeepMedic and our method respectively. The green and red crosses denote correctly and incorrectly localized lesion points.

TABLE VI

COMPARISON RESULTS (AVG ± STD DEV) OF CS LESION LOCALIZATION BETWEEN BASELINE METHODS AND OUR JOINTLY-OPTIMIZED MULTIMODAL CNNS BASED ON 5-FOLD CROSS-VALIDATION

	Sensi@0.1NLF	Sensi@1.0NLF
2D U-Net [50]	0.225 ± 0.067	0.833 ± 0.065
3D DeepMedic [51]	0.257 ± 0.104	0.772 ± 0.035
Ours	**0.637 ± 0.023**	**0.898 ± 0.021**

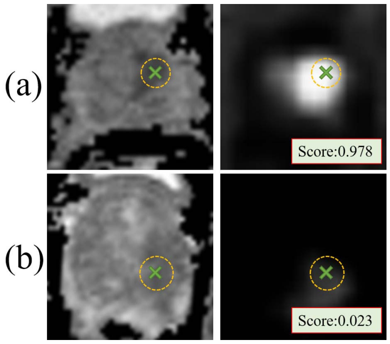

Fig. 11. Green cross signs denote the biopsy points. Yellow circles denote the 10mm neighborhood. Left: Original ADC images. Right: CRM_{CS} of the two images. Scores are calculated by averaging intensities of the CRM_{CS} within the yellow circles. (a) is an ADC image containing CS PCa and (b) is a nonCS cancerous ADC image.

cases. Thus the biopsy points reflecting locations of potential prostate cancer tissues are provided and participants of the challenge mainly focus on developing methods for classifying sub-regions around the provided biopsy points. For a fair comparison, we utilize the biopsy points provided by other participating methods in the challenge and calculate the probability of the corresponding lesion to be CS PCa by averaging the values of the CRM_{CS} within 10mm of the biopsy point based on the definition given in [3]. Specifically, CRM_{CS} is the cancer response map produced by our network with each pixel in CRM_{CS} indicating the probability of this pixel to be CS cancerous. Fig. 11 illustrates how we assign scores to lesions indicated by biopsy points based on CRM_{CS} generated by our network. We follow the same evaluation metric (i.e. AUC of the ROC) used in the PROSTATEx challenge. And due to the lack of ground-truth labels for the testing data, we perform a 5-fold cross validation on the training data of the challenge. Fig. 12 shows the average ROC curve of our proposed method and the AUC is 0.962 ± 0.0108. This AUC value is slightly higher than that of [52] which was 0.95 based on its validation set of the PROSTATEx training data. The method proposed in [52] achieved the second highest AUC in the PROSTATEx challenge.

F. Evaluations on Three Fine-Grained CS Lesion Localization Tasks

In addition to the evaluation on CS vs. nonCS lesion localization, we provide the performances of our proposed CAD system on three fine-grained CS lesion localization tasks: *Task-1*: indolent (i.e. PCa with GS ≤ 6) vs. CS PCa, *Task-2*: normal/BPH vs. CS PCa, *Task-3*: and BPH vs. central gland only CS PCa. The data used in this work includes 183 normal/BPH patients (82 are BPH and 101 lack of specific labels whether BPH or normal), 43 indolent PCa patients and 134 CS PCa patients (19 are central gland only CS PCa patients). The evaluations are still performed based on 5-fold cross-validations, while the evaluation metric NLF is re-defined as false positives per patient since there is no normal/BPH patient in task of indolent vs. CS PCa. Comparison results are shown in Fig. 13.

For the CS lesion localization of indolent vs. CS PCa, our CAD system achieves slightly worse performance (sensitivity of 0.389±0.155 at 0.1NLF and 0.805±0.035 at 1NLF) among all three tasks, which means that our CAD has limitations that wrongly localizing false positives for some indolent cases. Although distinguishing BPH from central gland only CS PCa seems like a difficult task since BPH is often in the

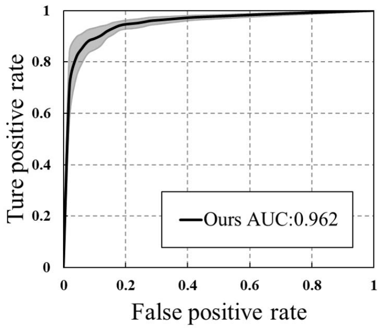

Fig. 12. Average ROC curves of our proposed system evaluated on the PROSTATEx training set based on 5-fold cross-validation. 95% confidence interval is shown as transparent area around the mean curve.

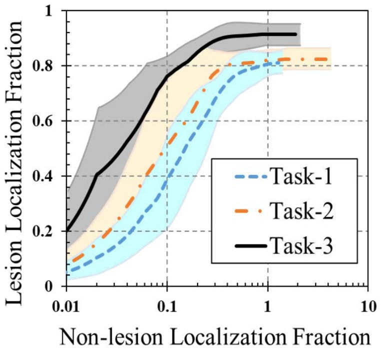

Fig. 13. Average FROC curves of our proposed CAD system for three different CS lesion localization tasks, i.e. indolent vs. CS PCa (Task-1), normal/BPH vs. CS PCa (Task-2), and BPH vs. central gland only CS PCa (Task-3). 95% confidence intervals are shown as transparent areas around the mean curves.

central gland, our CAD system achieves an extremely good performance (sensitivity of 0.763 ± 0.067 at 0.1NLF and 0.915 ± 0.036 at 1NLF).

V. Conclusion and Future Work

In this study, we present an end-to-end neural network for automatically detecting CS PCa in mp-MRI images. By concatenating TDN and multimodal CNN, the proposed neural network can jointly optimize all steps including multimodal registration, prostate detection, CS vs. nonCS classification and CRM generation, yielding a robust fault-tolerant CAD system which can accurately detect CS PCa. It also can be trained in a weakly-supervised manner by providing only image-level labels indicating the presence/absence of CS PCa without exact priors of lesions' locations. Those weakly-supervised annotations are much easier to obtain than manual delineation of PCa lesions which is required for most of existing CAD systems. Additionally, during the testing phase our network requires little manual configurations for setting parameters of the algorithms. As a result, our system is easy to use for radiologists/clinicians who have little knowledge about data preparation and algorithms. Experimental results on 246 patients' mp-MRI data demonstrated that our system's superior performance to that using individually optimized steps. Our future work includes extending the 2D network to 3D to better address the 3D deformation problem.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA, Cancer J. Clin.*, vol. 65, no. 1, pp. 5–29, Jan./Feb. 2015.

[2] A. Stangelberger, M. Waldert, and B. Djavan, "Prostate cancer in elderly men," *Rev. Urol.*, vol. 10, no. 2, pp. 111–119, 2008.

[3] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1083–1092, May 2014.

[4] Y. Peng *et al.*, "Quantitative analysis of multiparametric prostate MR images: Differentiation between prostate cancer and normal tissue and correlation with Gleason score—A computer-aided diagnosis development study," *Radiology*, vol. 267, no. 3, pp. 787–796, 2013.

[5] Y. Artan *et al.*, "Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2444–2455, Sep. 2010.

[6] D. Fehr *et al.*, "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 46, pp. E6265–E6273, 2015.

[7] G. Lemaitre, "Computer-aided diagnosis for prostate cancer using multi-parametric magnetic resonance imaging," Ph.D. dissertation, Dept. Comput., Univ. Burgundy, Dijon, France, 2016.

[8] G. J. S. Litjens, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman, "Automated computer-aided detection of prostate cancer in MR images: From a whole-organ to a zone-based approach," *Proc. SPIE*, vol. 8315, p. 83150G, Feb. 2012.

[9] G. J. S. Litjens, P. C. Vos, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman, "Automatic computer aided detection of abnormalities in multi-parametric prostate MRI," *Proc. SPIE*, vol. 7963, p. 79630T, Mar. 2011.

[10] P. Liu *et al.*, "A prostate cancer computer-aided diagnosis system using multimodal magnetic resonance imaging and targeted biopsy labels," *Proc. SPIE*, vol. 8670, p. 86701G, Feb. 2013.

[11] E. Niaf, O. Rouvière, F. Mège-Lechevallier, F. Bratan, and C. Lartizien, "Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI," *Phys. Med. Biol.*, vol. 57, no. 12, p. 3833, 2012.

[12] P. Tiwari, J. Kurhanewicz, and A. Madabhushi, "Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS," *Med. Image Anal.*, vol. 17, no. 2, pp. 219–235, Feb. 2013.

[13] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review," *Comput. Biol. Med.*, vol. 60, pp. 8–31, May 2015.

[14] S. Wang, K. Burtt, B. Turkbey, P. Choyke, and R. M. Summers, "Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research," *BioMed. Res. Int.*, vol. 2014, Dec. 2014, Art. no. 789561.

[15] V. Giannini *et al.*, "A fully automatic computer aided diagnosis system for peripheral zone prostate cancer detection using multi-parametric magnetic resonance imaging," *Comput. Med. Imag. Graph.*, vol. 46, pp. 219–226, Dec. 2015.

[16] V. Giannini *et al.*, "A prostate CAD system based on multiparametric analysis of DCE T1-w, and DW automatically registered images," *Proc. SPIE*, vol. 8670, p. 86703E, Feb. 2013.

[17] P. C. Vos, T. Hambrock, J. O. Barenstz, and H. J. Huisman, "Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI," *Phys. Med. Biol.*, vol. 55, no. 6, p. 1719, 2010.

[18] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[19] L. Matulewicz *et al.*, "Anatomic segmentation improves prostate cancer detection with artificial neural networks analysis of ^{1}H magnetic resonance spectroscopic imaging," *J. Magn. Reson. Imag.*, vol. 40, no. 6, pp. 1414–1421, Dec. 2014.

[20] E. Niaf, O. Rouvière, and C. Lartizien, "Computer-aided diagnosis for prostate cancer detection in the peripheral zone via multisequence MRI," *Proc. SPIE*, vol. 7963, p. 79633P, Mar. 2011.

[21] S. Ozer *et al.*, "Prostate cancer localization with multispectral MRI based on relevance vector machines," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro (ISBI)*, Jun./Jul. 2009, pp. 73–76.

[22] P. Puech, N. Betrouni, N. Makni, A.-S. Dewalle, A. Villers, and L. Lemaitre, "Computer-assisted diagnosis of prostate cancer using DCE-MRI data: Design, implementation and preliminary results," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 4, no. 1, pp. 1–10, Jan. 2009.

[23] P. C. Vos, T. Hambrock, J. O. Barentsz, and H. J. Huisman, "Combining T2-weighted with dynamic MR images for computerized classification of prostate lesions," *Proc. SPIE*, vol. 6915, p. 69150W, Mar. 2008.

[24] P. C. Vos, T. Hambrock, C. A. Hulsbergen-Van de Kaa, J. J. Fütterer, J. O. Barentsz, and H. J. Huisman, "Computerized analysis of prostate lesions in the peripheral zone using dynamic contrast enhanced MRI," *Med. Phys.*, vol. 35, no. 3, pp. 888–899, Mar. 2008.

[25] S. Viswanath *et al.*, "Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol *in vivo* 3 Tesla MRI," *Proc. SPIE*, vol. 7260, p. 72603I, Feb. 2009.

[26] S. Klein, U. A. van der Heide, I. M. Lips, M. van Vulpen, M. Staring, and J. P. W. Pluim, "Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information," *Med. Phys.*, vol. 35, no. 4, pp. 1407–1417, Apr. 2008.

[27] E. Niaf, R. Flamary, O. Rouviere, C. Lartizien, and S. Canu, "Kernel-based learning from both qualitative and quantitative labels: Application to prostate cancer diagnosis based on multiparametric MR imaging," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 979–991, Mar. 2014.

[28] P. Tiwari, S. Viswanath, J. Kurhanewicz, A. Sridhar, and A. Madabhushi, "Multimodal wavelet embedding representation for data combination (MaWERiC): Integrating magnetic resonance imaging and spectroscopy for prostate cancer detection," *NMR Biomed.*, vol. 25, no. 4, pp. 607–619, Apr. 2012.

[29] P. Vos, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman, "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis," *Phys. Med. Biol.*, vol. 57, no. 6, p. 1527, 2012.

[30] Z. Xinran *et al.*, "Fine-tuned deep convolutional neural network for automatic detection of clinically significant prostate cancer with multi-parametric MRI," in *Proc. Int. Soc. Magn. Reson. Med.*, 2017.

[31] I. Chan *et al.*, "Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier," *Med. Phys.*, vol. 30, no. 9, pp. 2390–2398, Sep. 2003.

[32] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.

[33] M. Jaderberg, K. Simonyan, A. Zisserman, and K. kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[34] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4168–4176.

[35] X. Yang *et al.*, "Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI," *Med. Image Anal.*, vol. 42, pp. 212–227, Dec. 2017.

[36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[39] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2843–2851.

[40] P. Tang, X. Wang, Z. Huang, X. Bai, and W. Liu, "Deep patch learning for weakly supervised object classification and discovery," *Pattern Recognit.*, vol. 71, pp. 446–459, Nov. 2017.

[41] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[42] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, nos. 285–296, pp. 23–27, 1975.

[43] L. Geert, D. Oscar, B. Jelle, K. Nico, and H. Henkjan. (2017). *Prostatex challenge data. The Cancer Imaging Archive*. [Online]. Available: https://doi.org/10.7937/K9TCIA.2017.MURS5CL

[44] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013.

[45] D. P. Chakraborty, "A brief history of free-response receiver operating characteristic paradigm data analysis," *Acad. Radiol.*, vol. 20, no. 7, pp. 915–919, Jul. 2013.

[46] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, 2017.

[47] Y. Sun, J. Yuan, W. Qiu, M. Rajchl, C. Romagnoli, and A. Fenster, "Three-dimensional nonrigid MR-TRUS registration using dual optimization," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1085–1095, May 2015.

[48] X. Cao, J. Yang, Y. Gao, Y. Guo, G. Wu, and D. Shen, "Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis," *Med. Image Anal.*, vol. 41, pp. 18–31, Oct. 2017.

[49] S. Raghunathan, D. Stredney, P. Schmalbrock, and B. D. Clymer, "Image registration using rigid registration and maximization of mutual information," presented at the 13th Annu. Med. Meets Virtual Reality Conf. (MMVR), 2005.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2015, pp. 234–241.

[51] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2016.

[52] S. Liu, H. Zheng, Y. Feng, and W. Li, "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI," *Proc. SPIE*, vol. 10134, p. 1013428, Mar. 2017.